Exploiter des corpus annotés syntaxiquement pour observer le continuum entre arguments et circonstants

Cécile Fabre, Didier Bourigault

Exploiter des corpus annotés syntaxiquement pour observer le continuum entre arguments et circonstants

Cécile Fabre et Didier Bourigault

CLLE (UMR 5263)
Université de Toulouse et CNRS

ABSTRACT

Dans cet article, nous proposons une méthode qui permet de mesurer le degré d'autonomie que manifestent les compléments prépositionnels vis-à-vis du verbe dans un corpus, de manière à tester l'hypothèse couramment admise d'un continuum entre arguments et circonstants et d'en étudier les manifestations. L'exploitation de corpus annotés catégoriellement et syntaxiquement et la mise au point de méthodes de quantification nous permettent de sonder ce continuum en divers points. La méthode met alors au jour des positions médianes, dont nous montrons qu'elles peuvent donner à voir des configurations récurrentes propres au corpus, au comportement intermédiaire entre arguments et circonstants prototypiques.

Exploiter des corpus annotés syntaxiquement pour observer le continuum entre arguments et circonstants

Cécile Fabre et Didier Bourigault

CLLE (UMR 5263)
Université de Toulouse et CNRS

I- INTRODUCTION

Aujourd'hui le travail sur des corpus de français s'est largement répandu. Nombreux sont les linguistes qui leur accordent désormais une place importante dans la phase de constitution de leurs observables. On constate néanmoins que ces linguistes atteignent rarement le même degré de familiarité avec les dispositifs techniques qui accompagnent les corpus – instruments d'annotation, d'exploration et d'analyse (Habert 2005). Cette « linguistique à l'instrument » que cherche à promouvoir Benoît Habert ne s'est pas encore imposée dans ses modalités les plus avancées. On s'en tient le plus souvent à l'association entre un corpus brut, c'est-à-dire réduit au seul texte, sans annotations linguistiques, et une interface de visualisation de type concordancier. De fait, la linguistique française s'est investie dans une démarche de constitution de corpus, impliquant collecte, transcription et annotation, mais elle tarde encore à tirer pleinement parti des corpus annotés.

Dans cet article, nous montrons l'intérêt de travailler en aval d'un étiquetage catégoriel et syntaxique pour calculer certaines propriétés linguistiques invisibles « à l'œil nu », à travers une expérience portant sur la distinction entre groupes prépositionnels arguments et circonstants. Sur cette question mille fois débattue, le sentiment est que la linguistique est arrivée à un point difficile à dépasser : les critères basés sur le jugement introspectif sont nombreux mais aucun n'est concluant, et un consensus s'est dégagé pour considérer qu'il existe un continuum entre ces deux catégories dont les pôles extrêmes seraient les seuls caractérisables. Nous proposons une méthode, basée sur un corpus annoté syntaxiquement, qui rend possible l'observation concrète de ce continuum.

La deuxième partie de l'article rappelle les objectifs d'un travail linguistique à partir de corpus annotés. Après avoir brièvement explicité, dans la troisième partie, l'objectif de caractérisation des types de complémentation, nous présentons, dans la quatrième partie, la méthode d'analyse, qui s'appuie sur l'extraction de groupes prépositionnels à partir d'un gros corpus annoté syntaxiquement et calcule une estimation de leur degré de dépendance par rapport au verbe à partir de propriétés observables

en corpus. Nous montrons pour finir, à travers quelques résultats, en quoi cette méthode permet effectivement de prolonger et renouveler la réflexion sur les propriétés de sélection des verbes et le statut des compléments prépositionnels.

II- LINGUISTIQUE DESCRIPTIVE SUR CORPUS ANNOTES

2.1. Pourquoi aller au-delà des corpus bruts ?

Pourquoi s'agirait-il d'aller au-delà des facilités qu'offre l'exploration de corpus nus à l'aide d'un concordancier ? Il est vrai que ce dispositif léger reste satisfaisant lorsqu'il s'agit de mener une étude sur un objet linguistique réductible à une forme lexicale, ou à quelques-unes (cf. par exemple le travail de Mondada (2005) sur *autrement*). L'exploration manuelle de concordances semble irremplaçable, la simplicité du corpus nu s'impose. C'est d'autant plus vrai lorsque le corpus collecté est rebelle à l'annotation automatique, du fait de sa nature : corpus oral, dont les particularités (répétitions, réparations) perturbent les outils d'étiquetage, corpus d'écrit spontané non révisé contenant fautes d'orthographe, simplifications graphiques ou grammaticales, etc.

Le recours aux corpus annotés s'avère cependant nécessaire dès qu'il s'agit de travailler sur des structures morpho-syntaxiques, et que l'objet d'étude ne peut pas être décrit par des patrons de fouille sur corpus bruts, comme l'explique Carruthers (2006: 251) :

> (…) in most cases, it is not possible to identify searchable items since they more often involve phenomena such as choice of word order or syntactic elements (for example, studies of detachment, negation, relatives, interrogation, use of pronouns), or verb morphology (for example, studies of tense or mood). This difficulty has led to the creation of various 'tagging' techniques, whereby phenomena are marked up by the researcher using codes and can subsequently be searched using specially designed programmes.

Lorsque les observables à dégager ne sont pas ancrés dans un matériau lexical spécifique, il est indispensable d'envisager les apports d'un dispositif qui combine corpus annotés et instruments d'analyse. Ainsi, Mela (2004), dans le cadre d'une expérience de repérage et d'extraction de gloses, fait la démonstration de l'apport de techniques d'étiquetage et d'extraction dans la mise au jour du phénomène linguistique qui l'intéresse. Elle conclut que les techniques informatiques renouvellent les conditions de l'observation linguistique : ce type de dispositif 'permet de vérifier l'opérationnalité des marques pressenties, de mesurer leur fréquence dans les différents types de textes mais aussi d' "apprendre" de nouvelles marques de glose à partir des textes' (Mela 2004 :76).

2.2. Disponibilité des corpus annotés

Le niveau d'annotation le plus pratiqué est l'étiquetage catégoriel, avec lemmatisation ; il peut être complété par une phase plus ou moins approfondie d'analyse syntaxique – parenthésage des constituants, identification de certaines relations de dépendance syntaxique. Le principal représentant des corpus de français annotés reste *Frantext*, qui est accessible depuis plusieurs années dans une version étiquetée[1] (plus de la moitié du corpus est actuellement accessible sous cette forme). Anne Abeillé et son équipe ont construit une banque d'arbres syntaxiques à l'Université de Paris 7 (Abeillé 2003). La maturité des techniques de TAL rend désormais possible l'annotation à la demande de corpus au niveau catégoriel (par exemple, avec un étiqueteur comme le Treetagger[2]), et, de plus en plus, syntaxique, puisque des analyseurs sont développés et deviennent utilisables (Paroubek 2007). Dans ce cas, il faut accepter le taux d'erreur lié à un étiquetage automatique : 'Il importe alors d'apprendre en linguistique à composer avec le caractère nécessairement imparfait des ressources (…) et à ne pas attendre des données annotées "pures", sans erreur, et ressortissant exactement aux distinctions de telle ou telle théorie.' (Habert 2004 :10).

Or, peu d'études linguistiques se sont jusqu'à présent emparé de ces données. A l'évidence, le traitement automatique des langues est à la fois pourvoyeur et utilisateur de corpus annotés, et la linguistique descriptive n'a guère de place dans ce cycle. Les linguistes disposent trop rarement des compétences techniques qui leur permettraient d'accéder à ces données, principalement du fait de la complexité de l'annotation à exploiter. Les instruments usuels de recherche et d'analyse de corpus sont loin d'être tous efficients (Pincemin 2004). Des compétences techniques spécifiques sont alors requises, à l'intersection de la linguistique et du TAL.

2.3. Combiner annotation et outils d'extraction et d'analyse

Leroy (2004), dans ses études sur l'antonomase du nom propre, est vite confrontée aux limites de la consultation de *Frantext* : faciles à formuler, les requêtes s'avèrent trop bruitées du fait de l'inadéquation du jeu d'étiquettes disponible (impossibilité en particulier de distinguer la nature du déterminant) ; par ailleurs, les caractéristiques de son objet d'étude l'amènent à souhaiter interroger des corpus plus adaptés. L'alternative est alors le développement d'un programme d'extraction d'antonomases et la mise au point de patrons de recherche fondés sur des étiquettes catégorielles plus conformes à son objet. L'investissement est important, mais l'auteur insiste sur deux apports : quantitatif d'abord, puisque 'la productivité de l'antonomaseur est nettement supérieure à celle du traditionnel glanage

[1] http://www.frantext.fr/categ.htm
[2] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

d'occurrences' ; qualitatif aussi, la mise au point de patrons de fouille par tests successifs contribuant à affiner la connaissance qu'elle a des propriétés formelles de son objet d'étude. C'est la démarche qu'ont adoptée également avant elle Tanguy et Rebeyrolle (2000) pour le repérage de contextes définitoires : cherchant à passer d'une description linguistique du phénomène à sa traduction en une représentation opérationnelle, ils ont élaboré des patrons de recherche combinant lemmes et catégories grammaticales, dont les performances ont été précisément évaluées, et qui ont fait l'objet de requêtes adressées à un concordancier sur corpus étiqueté. Ce type de démarche permet le passage à des entités de niveau supérieur pour dégager des structures mieux aptes à capter un phénomène linguistique donné, et facilite l'extraction d'occurrences pour permettre la quantification. La nécessité de produire des patrons de recherche opérationnels contribue à objectiver et à préciser la connaissance du phénomène à étudier.

Le travail que nous présentons ici, qui prolonge et complète une expérience précédente (Fabre et Frérot 2002), s'inscrit dans la continuité de ce type de travaux. Notre objectif est d'appliquer à la dichotomie argument/circonstant le critère d'opérationnalité que nous venons d'exposer : il s'agit de trouver des moyens d'extraire à partir de corpus des compléments susceptibles de relever de ces deux catégories, pour pouvoir les comparer et les étudier. Mais si nous partageons cet objectif d'opérationnaliser la description, nous proposons d'aller plus loin sur deux points : tout d'abord parce que nous nous basons sur des corpus analysés syntaxiquement – permettant la sélection d'entités caractérisées non seulement par leur catégorie mais aussi par leur position syntaxique ; ensuite parce que nous présentons une approche qui prolonge l'extraction de structures basée sur des patrons de recherche par des calculs visant à mettre au jour les propriétés des segments extraits.

III- LE PROBLEME DE LA DISTINCTION ARGUMENT/CIRCONSTANT

La dichotomie classique qui oppose arguments et circonstants a été abondamment étudiée et critiquée, sous des vocables divers : d'un côté, compléments de verbe, compléments essentiels, régis, sous-catégorisés, qui sont étroitement liés au verbe du point de vue syntaxique et sémantique ; de l'autre, compléments adjoints, ajouts, modifieurs, satellites, qui caractérisent secondairement les circonstances du procès décrit par le verbe. Il s'agit de deux façons bien différentes de se rattacher au verbe : les arguments sont 'les compléments qui s'attachent dans des combinaisons bien réglées', qui font partie de son 'schéma de construction' (Borillo 1990), et dont les 'propriétés sont régies de façon étroite par celui-ci' (Miller 1998) ; les circonstants sont ceux 'qui apparaissent dans son environnement mais qui ne dépendent pas directement de ses propriétés structurelles' (Borillo 1990),

dont ‘le statut est indépendant de celui du verbe’ (Miller 1998). Cette opposition ne se traduit cependant pas par des propriétés clairement marquées. Les synthèses de Miller (1998) et de Bonami (2000) rappellent la diversité des critères linguistiques mis en œuvre pour cerner cette opposition : obligatorité ou optionnalité syntaxique et sémantique, possibilité de déplacement en tête de phrase, itérabilité, aptitude à la passivation, à la topicalisation, degré de sélection par le verbe, etc. Les deux auteurs s’accordent sur l’absence de critère universel permettant d’opposer ces deux types de compléments, ainsi que sur les résultats contradictoires de certains critères. Selon Miller, on est confronté à un *continuum* et on peut tout au plus définir une classe d’arguments et de circonstants prototypiques, qui sont respectivement les objets directs et certains compléments locatifs ou temporels.

Pour la linguistique introspective, il est difficile d’envisager une suite, dès lors que l’on a montré qu’il existe une large zone rebelle aux critères binaires de caractérisation. La linguistique de corpus peut reprendre la question au point où la linguistique introspective l’a porté jusqu’ici, et poursuivre dans cette voie, à savoir: abandonner le principe d’une dichotomie pour adopter celui d’une échelle graduelle permettant d’évaluer le ‘degré de sélection’ entre un verbe et le groupe prépositionnel qui s’y rattache, ou, inversement, le ‘degré d’autonomie’ d’un GP par rapport au verbe de la proposition. Il s’agit donc de mettre à l’épreuve ce principe par l’analyse d’un grand nombre de configurations de dépendances attestées en corpus entre un verbe et un GP, situées entre les deux extrêmes du continuum. Se pose alors le problème de la collecte des observables. Comment constituer ce réservoir de cas à observer à partir d’un corpus ? Avec un corpus brut, à l’oeil nu, le linguiste de corpus devrait passer beaucoup de temps à lire un grand nombre d’énoncés pour constituter une base d’observables de taille satisfaisante, et sans outil il ne s’engagera pas dans une telle aventure. C’est dans ce contexte que s’impose le recours aux corpus annotés et à des instruments de calcul. Dans la section suivante, nous présentons une méthode qui, à partir d’un corpus annoté syntaxiquement, concrétise l’idée du continuum, et fournit un dispositif qui permet au linguiste d’isoler parmi les énoncés du corpus les configurations susceptibles d’être les plus pertinentes à analyser vis-à-vis du problème abordé.

IV- METHODE DE MESURE DU DEGRE D’AUTONOMIE DES GP D’UN CORPUS

4.1 L’annotation syntaxique.

Nous avons développé une méthode qui permet de calculer, pour un corpus donné *annoté syntaxiquement*, et pour un ensemble de groupes

prépositionnels (GP) issus de ce corpus, un coefficient mesurant le degré d'autonomie global de ces GP dans ce corpus. Le corpus sur lequel nous avons choisi de commencer à travailler (*Romans XXème*) est constitué d'un ensemble de 520 romans français du XXème siècle issu de la base *Frantext*[3]. La taille du corpus *Romans XXème* est d'environ 30 millions de mots.

L'annotation syntaxique est fournie par l'analyseur syntaxique Syntex (Bourigault & Fabre, 2000), (Bourigault, 2007), un analyseur en dépendance robuste, qui a été développé au sein du laboratoire CLLE-ERSS. Syntex est un analyseur incrémental organisé en « couches » (Abney, 1996 ; Aït-Moktar *et al.*, 2002), qui prend en entrée les résultats du Treetagger. L'analyse s'effectue par un enchaînement en cascade d'une suite de modules qui prennent en charge chacun une relation syntaxique.

Syntex est un analyseur déterministe : il fournit une seule analyse (parfois partielle) des phrases du corpus. En particulier, il résout lui-même les ambiguïtés de rattachement prépositionnel. Pour cela, il exploite un lexique qui associe à un ensemble de quelques dizaines de milliers de couples, constitués d'un mot (verbe, nom ou adjectif) et d'une préposition, un coefficient censé refléter la force d'association entre ce mot et cette préposition. Ce lexique a été construit entièrement automatiquement, à l'aide de l'analyseur Syntex lui-même, à partir d'un très gros corpus (10 années du journal *Le Monde*, 200 millions de mots) (Bourigault et Frérot, 2005). L'analyseur Syntex ne fait pas la distinction argument/circonstant. Dans la tâche de résolution des ambiguïtés prépositionnelles, Syntex cherche à trouver le bon recteur parmi une liste de recteurs potentiels qu'il a préalablement identifiés. Dans le cas où le recteur qu'il choisit est un verbe, il ne donne pas d'information sur le caractère argumental ou circonstanciel du groupe prépositionnel ainsi relié au verbe. Le taux de précision du rattachement prépositionnel se situe, selon les corpus, dans une fourchette de 85-90% (Bourigault et Frérot, 2005), (Bourigault, 2007)[4]. L'annotation syntaxique sur laquelle vont s'appuyer nos calculs n'est donc pas parfaite. Mais le travail sur gros corpus annotés requiert de savoir composer avec l'imparfait. Quand on calcule des tendances, des forces, à partir de très grands nombres d'occurrences, le quantitatif peut combler les lacunes du qualitatif. L'expérience que nous menons doit nous permettre de juger si l'annotation syntaxique automatique fournie par Syntex constitue un compromis satisfaisant entre les exigences de qualité et de quantité imposées par la méthode.

---

[3] Nous remercions Jean-Marie Pierrel, directeur de l'ATILF, de nous avoir fourni ce corpus de textes, dont certains sont encore sous droits, à des fins de recherche.

[4] L'analyseur Syntex a obtenu les meilleurs résultats lors de la campagne d'évaluation Easy des analyseurs syntaxiques du français (Paroubek *et al.* 2007).

4.2 La méthode

Notation

Une fois le corpus analysé et les GP rattachés aux verbes par Syntex, la méthode calcule un coefficient mesurant le degré d'autonomie global de ces GP. Nous faisons l'hypothèse que plus ce coefficient est élevé, plus le GP a tendance à être employé dans le corpus comme circonstant, avec une force d'association faible avec le verbe dont il dépend. Les GP sont caractérisés sous la forme de couples (*p,n*), où *p* est une préposition et *n* un nom. Le couple (*p,n*) est le représentant normalisé de tous les GP du corpus introduits par la préposition *p* et dont la tête nominale est *n*. Pour être plus précis, on distingue ces GP selon que le nom tête est déterminé ou non. Cette distinction est notée par un indice sur le nom (D ou 0) :

- le couple ($à,vitesse_D$) est le représentant normalisé de tous les GP du corpus introduits par le préposition *à* et dont la tête nominale est le nom *vitesse* quand celui-ci est déterminé : *à la vitesse de 50 km/h*, *à une vitesse très élevée*, *à cette vitesse*, etc.
- le couple ($à,vitesse_0$) est le représentant normalisé de tous les GP du corpus introduits par le préposition *à* et dont la tête nominale est le nom *vitesse* quand celui-ci n'est pas déterminé : *à vitesse raisonnable*, *à grande vitesse*, etc.

Extraction de triplets (*v,p,n*)

A partir de l'ensemble des phrases du corpus analysées syntaxiquement en dépendance, on extrait les triplets (*v,p,n*) des configurations dans lesquelles le nom *n* est régi par la préposition *p*, elle-même régie par le verbe *v*. Par exemple, à partir de l'analyse par Syntex de la phrase *Tournant la tête, il vit l'ombre courir entre les troncs à la vitesse d'un cheval au galop*, on extrait le triplet ($courir,à,vitesse_D$) ; à partir de la phrase *Les ruisseaux coulaient à vitesse différente pour étancher les soifs les plus diverses*, on extrait le triplet ($courir,à,vitesse_0$).

Pour la suite des calculs, on ne retient que les triplets dont le nombre d'occurrences dans le corpus est supérieur à un certain seuil *s*, fixé à 2 dans la présente étude. Le nombre total de triplets ainsi retenus à partir du corpus *Romans XXème* est de 94 843.

Calcul du degré de sélectivité des couples (*v,p*)

Dans une première étape, on calcule une force d'association globale entre un verbe et une préposition. Il existe un grand nombre de méthodes pour calculer, à partir de corpus, la force d'association entre unités lexicales (Manning et Schütze, 1999). Nous reprenons ici une méthode originale, exploitée dans l'analyseur Syntex, que nous avons déjà testée sur la problématique de la distinction argument/circonstant (Fabre et Frérot,

2002). L'idée de cette méthode est, concernant l'association entre un mot et une préposition, d'accorder une importance non pas tant à la fréquence de coccurrence des deux éléments qu'à la *diversité* des contextes nominaux rencontrés avec ce couple. On définit ainsi la *productivité* d'un couple (*v,p*) comme le nombre de régis nominaux *n différents* qui entrent dans la construction (*v,p,n*) (Bourigault et Fabre, 2000). Une productivité élevée est l'indice que le verbe fonctionne de manière régulière avec cette préposition, donnant lieu à des occurrences diverses. L'expérience montre en effet que, dans des corpus thématiques, la haute fréquence de certains syntagmes répétitifs vient biaiser la mesure d'association lexicale. La méthode proposée vise à limiter une telle surestimation et à privilégier les associations diversifiées. Dans le corpus *Romans XXème*, le couple (*dépendre,de*) a une productivité de 50 : on le retrouve avec 50 noms régis différents : $personne_0$ (18 fois), $jour_D$ (15 fois), $volonté_D$ (10), $circonstance_D$ (6), etc. Le couple (*écrire,sur*) a aussi une productivité de 50 : $papier_D$ (32), mur_D (30), $feuille_D$ (22), etc. Le verbe *dépendre* ne se construit quasiment qu'avec la préposition *de*, alors que le verbe *écrire* se construit avec un grand nombre de prépositions différentes (*avec*, *dans*, *pour*, etc.). Pour mesurer le *degré de sélection* d'un verbe pour une préposition, on calcule alors une *productivité relative* en divisant la productivité du couple (*w,p*) par la productivité *totale* du verbe, qui est la somme de ses productivités pour toutes les prépositions avec lesquelles il se construit (tableau 1). On constate ainsi dans le tableau 2 que le degré de sélectivité du verbe *dépendre* pour la préposition *de* est très élevé, alors que celui du verbe *écrire* pour la préposition *sur* est faible.

$$\text{prod}(v,p)=\text{Card}\{n \,/\, \text{f}(v,p,n)>= \text{s}\}$$
$$\text{prod}_T(v)= \sum_p \text{prod}(v,p)$$
$$\text{selec}(v,p)= \text{prod}(v,p)/\text{prod}_T(v)$$

Tableau 1 : Définition du degré de sélection d'un verbe *v* pour une préposition *p*

v	*p*	$\text{prod}_T(v)$	prod(*v,p*)	selec(*v,p*)
dépendre	de	51	50	0.98
écrire	à	325	126	0.39
écrire	dans	325	37	0.11
écrire	sur	325	50	0.15
mourir	à	266	39	0.15
mourir	dans	266	50	0.19
mourir	de	266	78	0.29

Tableau 2 : Degrés de sélectivité des verbes *dépendre*, *écrire* et *mourir* avec différentes prépositions (selec>0.1)

Calcul du degré d'autonomie des couples (*p,n*)

Dans une seconde étape, de façon symétrique à ce qui a été défini pour les couples (*v,p*), on définit la productivité d'un couple (*p,n*) comme le nombre de verbes *v* différents qui entrent dans la construction (*v,p,n*). L'hypothèse de base de la méthode est que l'aptitude d'un GP à s'associer à une grande diversité de verbes est un premier signe de son autonomie sémantique. Par exemple, le couple (*dans, rue*$_D$) se construit avec 259 verbes différents, alors que le couple (*dans, dossier*$_D$) apparaît avec seulement 6 verbes (*lire*, *fouiller*, *chercher*, *voir*, *trouver*, *ranger*). Cependant, ce critère accorde un poids identique à chaque verbe, que son degré de sélection pour la préposition soit élevé – par exemple, *plonger* ou *s'enfoncer* vis-à-vis du couple (*dans, rue*$_D$), ou qu'il soit faible – par exemple *se battre*, ou *rêver*. Pour affiner la mesure, on définit une productivité pondérée $prod_P(p,n)$ comme la somme des degrés de sélection des verbes avec lesquels se construit le couple (*p,n*). A productivité égale, une productivité pondérée élevée est un indice négatif quant à la capacité du GP à être autonome. On peut ainsi faire l'hypothèse que plus le rapport entre la productivité pondérée et la productivité est élevé – c'est-à-dire plus la proportion de verbes sélectifs se construisant avec le GP est élevée – moins on est assuré de l'autonomie du GP. C'est pourquoi on définit la mesure d'autonomie du GP comme l'écart à 1 de ce ratio, dont la valeur est comprise entre 0 et 1 (tableau 3).

$$prod(p,n)=\text{Card}\{v \,/\, f(v,p,n)>= s\}$$

$$prod_P(p,n)= \sum_{\{v \,/\, f(v,p,n)>= s\}} selec(v,p)$$

$$auton(p,n)= 1 - (prod_P(p,n) \,/\, prod(p,n)$$

Tableau 3 : Définition du degré d'autonomie d'un couple(*p,n*)

Nous illustrons la mesure d'autonomie avec plusieurs exemples. Les couples (*à, frère*$_D$) et (*à, cuisine*$_D$) ont tous les deux une productivité de 44 : ils se construisent avec 44 verbes différents. Le coefficient d'autonomie du couple (*à, frère*$_D$) est de 0.64, alors que celui du couple (*à, cuisine*$_D$) est de 0.38 :

- pour le couple (*à,cuisine*$_D$), il n'y a que 2 verbes dont le degré de sélection pour la préposition *à* est supérieur à 0.75 : *attacher* et *penser* ; alors qu'il y a 20 verbes dont le degré de sélection pour la préposition *à* est inférieur à 0.25 : *trouver*, *retrouver*, *chercher*, *attendre*, *voir*, etc. Le degré d'autonomie du couple (*à,cuisine*$_D$) est élevé.

- pour le couple ($à,frère_D$), il y a 20 verbes dont le degré de sélection pour la préposition *à* est supérieur à 0.75 : *emprunter*, *ressembler*, *appartenir*, *recommander*, *s'adresser*, etc. ; alors qu'il n'y a que 4 verbes dont le degré de sélection pour la préposition *à* est inférieur à 0.25 : *passer*, *faire*, *souffler*, *parler*. Le degré d'autonomie du couple ($à,frère_D$) est faible.

Cette méthode ordonne ainsi les couples (*p,n*) du corpus, concrétisant l'hypothèse du continuum. Dans la dernière partie, nous allons présenter les résultats en sondant différents endroits du continuum.

V- RESULTATS

5.1. Constats préliminaires

Comportement général des prépositions

Nous nous intéressons tout d'abord à des résultats globaux concernant la propension de chacune des principales prépositions à donner lieu à des GP plutôt autonomes ou plutôt dépendants du verbe. La figure 1 donne l'histogramme des degrés d'autonomie des GP pour les principales prépositions. On constate que seules les prépositions *à* et *de* occupent tout le spectre des valeurs possibles, et qu'elles sont moins enclines que les autres prépositions à introduire des GP à fort degré d'autonomie. Les autres prépositions présentent quant à elles des tendances plus ou moins marquées à l'autonomie, les prépositions *en* et *avec* étant clairement situées toutes deux dans la zone de plus forte autonomie, alors que *dans* et *sur* offrent une image plus contrastée. L'opposition usuelle entre prépositions vides ou incolores et prépositions circonstancielles se trouve ici concrétisée, et ceci constitue un premier signe encourageant sur la pertinence linguistique de la méthode. Nous allons entrer dans le détail de ces fonctionnements, en commençant par rendre compte des situations les plus nettes, où les GP se situent aux deux pôles du continuum.

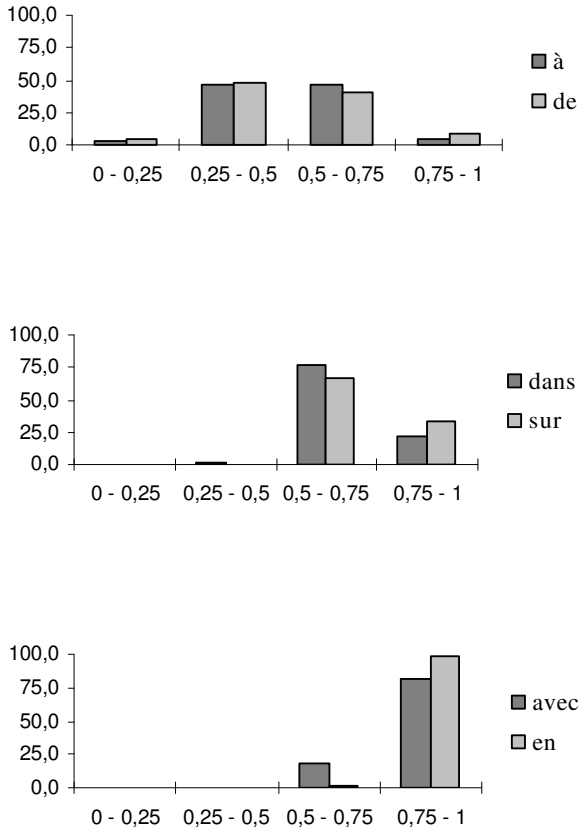

Fig. 1 : Histogramme des degrés d'autonomie des couples (*p,n*) pour les préposition *à*, *de*, *dans*, *sur*, *avec*, *en*

Autonomie maximale : conjonctions et adverbiaux

Dans les valeurs les plus hautes de la mesure d'autonomie des couples, nous trouvons un ensemble de couples (*p,n*) qui ne relèvent pas de la distinction arguments/circonstants. L'analyse en corpus des GP qu'ils représentent montre qu'ils sont issus de locutions prépositionnelles qui n'ont pas été reconnues comme telles par l'analyseur Syntex : par exemple, les couples ($à,approche_D$) et ($dans,espoir_D$), pour lesquels notre méthode calcule un très fort degré d'autonomie, sont issus respectivement des locutions prépositionelles *à l'approche de* et *dans l'espoir de*. Des couples très autonomes peuvent aussi être issus de groupes prépositionnels dont le degré de figement est tel qu'ils sont habituellement décrits comme des locutions adverbiales : les couples ($à,enjambée_0$), ($de,ordinaire_0$) et ($dans,circonstances_D$) sont en fait issus respectivement des locutions *à grandes enjambées*, *d'ordinaire*, *dans ces circonstances*. On rencontre ici

des contraintes liées à l'utilisation d'un corpus annoté syntaxiquement par un outil informatique, sans révision manuelle. Si les locutions prépositionnelles ou adverbiales énumérées ci-dessus avaient été considérées comme des unités (*tokens*) par l'analyseur, les couples incriminés n'auraient pas été extraits par la méthode. On peut en tirer deux conclusions : (i) le fait même que ce type de groupes intrus soient classés en tête de liste par la méthode constitue plutôt un gage de la pertinence de celle-ci ; (ii) il faut s'accommoder de ce type de découverte, et savoir mettre de côté les résultats qui n'entrent pas dans le périmètre de l'étude. C'est ce que nous faisons dans la suite de cette section, en réservant nos investigations aux couples (*p,n*) représentant de réels groupes prépositionnels.

5.2. Les deux pôles du continuum

Le tableau 4 présente, pour les prépositions *à*, *de*, *dans*, des exemples de couples appartenant au pôle positif du continuum (autonomie maximale). On a affaire à des couples renvoyant à des séquences syntaxiquement et sémantiquement autonomes, porteuses d'informations circonstancielles relatives au temps (*connaître dans sa jeunesse*), à la manière (*déclarer d'un ton Adj*), à l'espace (*découvrir à l'horizon*). Ils contrastent fortement avec les couples relevés à l'autre extrémité du continuum, illustrés dans le tableau 5. Ces couples renvoient quant à eux à des GP fortement dépendants du verbe : *renoncer à l'effort, tapisser de feuilles, se perdre dans la contemplation*. Ces GP sont logiquement difficiles à caractériser en termes de contribution sémantique : contrairement aux séquences autonomes, leur sémantique est déterminée par le verbe dont ils dépendent. On constate néanmoins une plus grande homogénéité sémantique dans le cas de *dans*, du fait de l'apport sémantique propre de la préposition. En effet, les GP en *dans* situés dans ce pôle négatif correspondent à des séquences marquant principalement l'intériorité dans un espace clos – concret ou abstrait, comme le montrent les verbes associés : *plonger*, *se jeter*, *pénétrer*, *entrer*…

Le contraste entre les deux pôles semble net pour une frange des groupes prépositionnels traités. Mais le principe d'une franche dichotomie cède très vite le pas : ainsi, si les couples (*dans*,*n*) peuvent se répartir sur deux pôles, les valeurs d'autonomie basses restent relativement élevées (pratiquement 0.5) par rapport aux prépositions *à* et *de*, suggérant un degré de sélection moindre de la part des verbes concernés. On quitte ainsi rapidement des valeurs extrêmes pour rejoindre des positions intermédiaires dont nous allons étudier la signification.

à		de		dans	
(*à*,*horizon*$_D$)	0.79	(*de*,*ton*$_D$)	0.87	(*dans*,*style*$_D$)	0.91
(*à*,*vitesse*$_D$)	0.77	(*de*,*trait*$_D$)	0.82	(*dans*,*langue*$_D$)	0.89

(*à*,*lueur*$_D$)	0.75	(*de*,*manière*$_D$)	0.82	(*dans*,*an*$_D$)	0.86
(*à*,*crayon*$_D$)	0.75	(*de*,*voix*$_D$)	0.81	(*dans*,*moment*$_D$)	0.85
(*à*,*surface*$_D$)	0.75	(*de*,*façon*$_0$)	0.78	(*dans*,*jeunesse*$_D$)	0.85

Tableau 4 : Exemples de couples (*p*,*n*) à *forte* autonomie avec les prépositions *à*, *de* et *dans*

à		de		dans	
(*à*,*effort*$_D$)	0.12	(*de*,*feuille*$_O$)	0.09	(*dans*,*contemplation*$_D$)	0.38
(*à*,*volonté*$_D$)	0.14	(*de*,*nuage*$_D$)	0.14	(*dans*,*flot*$_D$)	0.44
(*à*,*exigence*$_D$)	0.15	(*de*,*voile*$_D$)	0.18	(*dans*,*lecture*$_D$)	0.45
(*à*,*émotion*$_D$)	0.18	(*de*,*droit*$_D$)	0.18	(*dans*,*gouffre*$_D$)	0.46
(*à*,*influence*$_D$)	0.19	(*de*,*spectacle*$_D$)	0.19	(*dans*,*sol*$_D$)	0.47

Tableau 5 : Exemples de couples (*p*,*n*) à *faible* autonomie avec les prépositions *à*, *de* et *dans*

4.3. Positions intermédiaires

L'analyse des couples de degré d'autonomie moyen (entre 0.4 et 0.7) fait apparaître deux types de couples qui s'opposent quant au profil de répartition du degré de sélection des verbes avec lesquels ils se construisent.

Les couples du premier type présentent un profil nettement contrasté, une partie des verbes possédant un fort degré de sélection et la majorité des autres un degré de sélection faible. Ces couples sont amenés à jouer dans le corpus tantôt le rôle de circonstant, tantôt le rôle d'argument, selon les verbes qui les régissent. C'est le cas par exemple du couple (*à*,*portière*$_D$), dont l'histogramme est présenté sur la figure 2 : 1/3 environ des verbes avec lesquels on le trouve sont fortement sélectifs pour la préposition *à* (*s'adosser, s'accouder, accrocher*), tous les autres ayant au contraire un taux de sélection bas, voire très bas (*se précipiter*, *regarder*, *apparaître*). C'est le cas également de nombreux compléments marquant le lieu, envisagés soit comme destination d'un procès télique (*pénétrer dans le grenier*), soit comme simple localisation (*travailler dans le grenier*). Ces cas gagneraient sans doute à être étudiés dans le cadre d'une réflexion sur la sémantique du nom (et des différentes facettes qu'il instancie dans le corpus), mais il semble qu'on n'ait pas véritablement affaire ici à des cas intéressants du point de vue de la réflexion sur le continuum qui caractérise la nature de la complémentation : en effet, ces cas médians ne sont que le résultat de la rencontre entre des types de fonctionnement assez nettement tranchés – tantôt argument, tantôt circonstant.

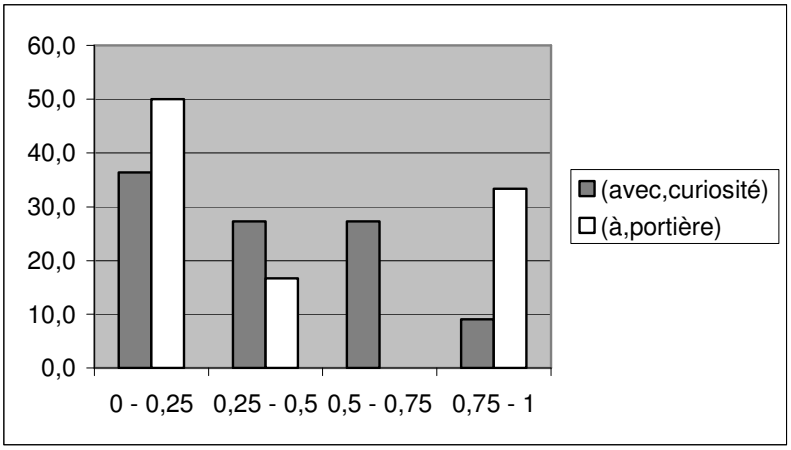

Fig. 2 : Histogramme des degrés de sélection des verbes pour les couples (*avec*, *curiosité*$_D$) et (*à*, *portière*$_D$)

Les couples du second type présentent des profils plus homogènes : une large proportion des verbes avec lesquels ils se construisent affichent un degré de sélection moyen. Ces couples correspondent donc à des GP qui, dans une grande partie de leurs occurrences, présentent une autonomie moyenne. Le couple (*avec*,*curiosité*$_D$), dont le profil est représenté sur la figure 2, relève de ce cas de figure : plus de la moitié des verbes auxquels il se rattache a un taux de sélection moyen. Cet exemple illustre un cas véritablement intermédiaire. Le tableau 6 dresse la liste des verbes auxquels on le trouve associé, avec leur degré de sélection pour la préposition *avec*.

verbe	degré de sélection
dévisager	0.82
contempler	0.67
examiner	0.6
écouter	0.6
regarder	0.37
observer	0.36
considérer	0.29
suivre	0.11
fixer	0.1
attendre	0.09

se pencher	0.05

Tableau 6 : liste des verbes se construisant avec le couple (*avec*,*curiosité*$_D$) avec leur degré de sélection pour la préposition *avec*

On constate que seul le verbe *dévisager* a un taux de sélection élevé avec la préposition *avec* (*dévisager avec : attention*$_D$, *curiosité*$_D$, ét*onnement*$_D$, *impertinence*$_D$, *insistance*$_D$...). Les autres verbes ont un taux de sélection plus faible, mais on a affaire à une gamme de verbes sémantiquement proches de *dévisager* (*contempler*, *examiner*, *regarder*, *observer*, *considérer*, *fixer*). Cet exemple illustre ainsi deux intérêts de la méthode :

1) Elle met au jour des cas d'associations récurrentes caractéristiques du corpus. Ainsi, si on ne peut pas considérer que le complément en *avec* est argument du verbe *dévisager*, on constate que cette association est néanmoins très régulière : si le verbe *dévisager* admet un complément – ce qui n'est bien sûr pas toujours le cas, ce verbe étant d'ordinaire un simple transitif – c'est quasiment toujours un complément en *avec*.
2) Elle permet de repérer des classes de verbes qui favorisent l'expression de certains compléments. Comme les observations que nous venons de faire pour (*avec*,*curiosité*$_D$) s'étendent aussi à d'autres compléments en *avec* (*avec stupeur*$_D$, *avec attention*$_D$), on peut déterminer un *pattern* sémantique bien établi dans le corpus : VERBE D'OBSERVATION + COMPLEMENT DE MANIERE.

L'observation des positions médianes permet donc de repérer des comportements véritablement intermédiaires entre circonstants et arguments. On se situe au-delà de la simple relation circonstancielle puisqu'on met en évidence des associations sémantiques fortes entre certains types de GP et certains verbes. Mais on reste en-deçà de la relation argumentale puisque le complément n'est pas requis, n'apparaît pas toujours, et les tests montreraient qu'il possède d'autres propriétés des circonstants – le déplacement et l'extraction sont possibles.

VI- CONCLUSION

Dans cette expérience, nous avons cherché à vérifier empiriquement l'hypothèse d'un continuum entre compléments argumentaux et complément circonstanciels. Notre méthode mesure le degré d'autonomie des compléments prépositionnels vis-à-vis du verbe dont ils dépendent, grâce à des tests élaborés sur un grand corpus analysé syntaxiquement. Cette approche donne les moyens de dépasser l'étude des seuls exemplaires prototypiques clairement argumentaux ou circonstanciels, pour mettre au jour des situations intermédiaires. Les premiers exemples que nous avons

commentés montrent en effet qu'en identifiant de manière automatique des compléments prépositionnels dont le degré d'autonomie est moyen on peut repérer dans le corpus des combinaisons récurrentes qui concernent des adjoints dont l'expression est favorisée par la sémantique du verbe.

L'observation de ces positions médianes devra être approfondie, en particulier en confrontant les résultats de plusieurs corpus. Il sera en effet intéressant de vérifier la variabilité du degré de sélection des verbes et d'autonomie des GP dans différents corpus, notre hypothèse étant que les associations récurrentes sont fortement dépendantes du genre du corpus considéré.

De façon plus générale, nous avons voulu montrer à travers cette expérience que l'exploitation de grands corpus annotés syntaxiquement ouvre la voie à des études empiriques où des hypothèses linguistiques sont testées grâce au recours à des instruments adéquats de quantification.

REFERENCES

Abeillé, A. (ed.) (2003). *Building and Using Parsed Corpora*. Dordrecht: Kluwer.

Abney, S. (1990). Parsing by chunks. In: R. Berwick, S. Abney et D. Tenny (eds), *Principle-Based Parsing*. Dordrecht: Kluwer, pp. 257-278.

Aït-Mokhtar S., Chanod J.-P. et Roux C. (2002). Robustness beyond shallowness: incremental deep parsing. *Natural Language Engineering* 8: 121-144.

Bonami, O. (2000). Les constructions du verbe : le cas des groupes prépositionnels argumentaux : analyse syntaxique, sémantique et lexicale. Lille : A.N.R.T., Université de Lille 3.

Borillo, A. (1990). A propos de la localisation spatiale. *Langue française*, 86: 75-84.

Bourigault, D. et Fabre, C. (2000). Approche linguistique pour l'analyse syntaxique de corpus. *Cahiers de Grammaire*, 25: 131-151.

Bourigault, D. et Frérot C. (2005). Acquisition et évaluation sur corpus de propriétés de sous-catégorisation syntaxique. In: *Actes de la 12ème conférence francophone sur le Traitement Automatique des Langues Naturelles* (TALN). Dourdan, pp. 373-382.

Bourigault, D. (2007). *Un analyseur syntaxique opérationnel : Syntex*. Mémoire d'Habilitation à Diriger les Recherches, Université de Toulouse-Le Mirail.

Carruthers, J. (2006). État présent: The syntax of oral French. *French Studies : A Quarterly Review*, 60: 251-260.

Fabre, C. et Frérot, C. (2002). Groupes prépositionnels arguments ou circonstants: vers un repérage automatique en corpus. In: *Actes de la 9ème Conférence francophone sur le Traitement Automatique des Langues Naturelles* (TALN). Nancy, pp. 215-224.

Habert, B. (2004). Outiller la linguistique: de l'emprunt de techniques aux rencontres de savoirs. Dossier *Linguistique et informatique: nouveaux défis*. *Revue de linguistique appliquée*, 9: 5-24.

Habert, B. (2005). *Instruments et ressources électroniques pour le français*. Gap/Paris: Ophrys.

Leroy, S. (2004). Extraire sur patrons : allers et retours entre analyse linguistique et repérage automatique, *Revue française de linguistique appliquée*, 9: 25-43.

Manning, C. D. et Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts: MIT Press.

Mela, A. (2004). Linguistes et 'TAListes' peuvent coopérer : repérage et analyse des gloses. Dossier *Linguistique et informatique: nouveaux défis*. *Revue Française de Linguistique Appliquée,* 9: 63-82.

Miller P. (1998). Compléments et circonstants: une distinction syntaxique ou sémantique? In: J.-C. Souesme (ed.), *Cycnos* 15. Actes du *37ème Congrès de la SAES* (Société des Anglicistes de l'Enseignement Supérieur). Nice: Presses Universitaires de Nice, pp. 91-103.

Mondada, L. (2005). Utilisation de corpus pour l'évaluation d'hypothèses linguistiques: étude de *autrement*. In: A. Condamines (ed), *Sémantique et corpus*. Paris: Hermès, Lavoisier, pp. 109-145.

Paroubek, P., Vilnat, A., Robba, I. et Ayache, C. (2007). Les résultats de la campagne EASY d'évaluation des analyseurs syntaxiques du français. In: *Actes de la 14ème conférence francophone sur le Traitement Automatique des Langues Naturelles* (TALN). Toulouse, pp. 243-252.

Pincemin, B. (2004). Lexicométrie sur corpus étiquetés. In G. Purnelle et al. (eds.), *Le poids des mots*. Actes des 7èmes journées internationales d'analyse statistique des données textuelles (JADT), vol. II, pp. 865-873.

Tanguy, L. et Rebeyrolle, J. (2000). Repérage automatique de structures linguistiques en corpus : le cas des énoncés définitoires. *Cahiers de Grammaire*, 25: 153-174.